

# A Retrieval-Based Dialogue System Utilizing Utterance and Context Embeddings

Citation for published version (APA):

Bartl, A., & Spanakis, G. (2017). A Retrieval-Based Dialogue System Utilizing Utterance and Context Embeddings. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1120-1125). IEEE. <https://doi.org/10.1109/ICMLA.2017.00011>

**Document status and date:**

Published: 01/12/2017

**DOI:**

[10.1109/ICMLA.2017.00011](https://doi.org/10.1109/ICMLA.2017.00011)

**Document Version:**

Publisher's PDF, also known as Version of record

**Document license:**

Taverne

**Please check the document version of this publication:**

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.umlib.nl/taverne-license](http://www.umlib.nl/taverne-license)

**Take down policy**

If you believe that this document breaches copyright please contact us at:

[repository@maastrichtuniversity.nl](mailto:repository@maastrichtuniversity.nl)

providing details and we will investigate your claim.

# A retrieval-based dialogue system utilizing utterance and context embeddings

Alexander Bartl and Gerasimos Spanakis  
Department of Data Science and Knowledge Engineering  
Maastricht University  
6200MD, Netherlands

Email: a.bartl@student.maastrichtuniversity.nl, jerry.spanakis@maastrichtuniversity.nl

**Abstract**—Finding semantically rich and computer-understandable representations for textual dialogues, utterances and words is crucial for dialogue systems (or conversational agents), as their performance mostly depends on understanding the context of conversations. In recent research approaches, responses have been generated utilizing a decoder architecture, given the distributed vector representation (embedding) of the current conversation. In this paper, the utilization of embeddings for answer retrieval is explored by using Locality-Sensitive Hashing Forest (LSH Forest), an Approximate Nearest Neighbor (ANN) model, to find similar conversations in a corpus and rank possible candidates. Experimental results on the well-known Ubuntu Corpus (in English) and a customer service chat dataset (in Dutch) show that, in combination with a candidate selection method, retrieval-based approaches outperform generative ones and reveal promising future research directions towards the usability of such a system.

**Index Terms**—Dialogue Systems, Deep Learning, Information Retrieval

## I. INTRODUCTION

Text-only based Dialogue systems, also called Conversational Agents, Chatbots or Chatterbots, have become very popular in the research community and for large companies. The reason for the rise in popularity lies in the fact that their ability to interact intelligently with humans has improved significantly due to advancements in hardware technologies and Artificial Intelligence.

One of the latest effective approaches [1] is to represent words, phrases, or even complete dialogues as fixed-length vectors of floating point numbers, also called embeddings (or distributed representations or feature vectors). The Hierarchical Recurrent Encoder-Decoder (HRED) [2] and its successors [3], (as well as similar related models) are specifically designed to encode the meaning of textual conversations regarding the special structure that originates from multiple turn-taking speakers.

In our approach, a context embedding, a vector encoding the meaning of a conversation up to a certain time step  $t$ , encoded by the HRED model, serves as input to the decoder component to generate a textual answer. We explore the performance of a retrieval-based model that uses the utterance- and context-embeddings, previously generated by the HRED model, to find similar conversations and rank possible candidate answers. We argue that a retrieval-based approach, based on embeddings, can outperform the generative approach, as the retrieval of

similar conversations is less dependent on high quality embeddings and less susceptible to poorly trained embeddings.

The rest of the paper is organized as follows: We first give an outline of current research around dialogue systems. The proposed pipeline is discussed in Section III. Experimental setup (datasets, evaluation metrics, models implemented and compared) as well as the results are discussed in Section IV. Finally, we conclude the paper by summarizing the main findings and outline future work.

## II. RELATED WORK

The purpose of Dialogue Systems (DS), often also termed Conversational Agents (CA) or Chatterbots, is to converse with humans to provide information, help in decision making, perform administrative services, or just for the sake of entertainment [4].

One of the early simple approaches to dialogue systems was to simply spot certain key-words or combinations of them (like in the case of script-based chatbots). With advances in machine learning, the development ranges from statistical modelling of language [5], semantic parsing [6], skip-gram models [1], and others, to approaches utilizing deep neural architectures [7]. Neural networks have also been used to improve language generation [8]. Recently, Dialogue Managers have made similar advances towards automated solutions, with a focus on reinforcement learning [9], generating policies of how to interact with humans, based on some state representation.

With the rise of Deep Learning (DL) in recent years [10] and an increasing company interest in chatterbots, end-to-end Dialogue Systems, such as deep Recurrent Neural Networks, constituting all modules in one model [2], have become one of the major research topics for Dialogue Systems. Such a system would generate an answer end-to-end from raw user input.

Our proposed pipeline is a combination of a generative- and retrieval-based approach. An encoder model, such as the HRED model (or could be one of its more advanced variations) is trained end-to-end on a textual corpus, using an objective function that is based on how capable the model is of generating the answers in the training set. After the training however, the decoder component of the HRED model is not used to generate answers. Instead, we argue that a retrieval-based approach taking over the language

part performs equally or better in both general and specific domains.

### III. MODEL DESCRIPTION

The proposed model can be split up into three individual components. The first component, the encoder, utilizes the HRED model to encode raw conversations into embeddings containing the actual meaning. The second component, a retrieval-based approach using an Approximate Nearest Neighbor (ANN) model, is responsible for retrieving similar conversations from a database of embedding- and raw-text-tuples. Given the context of an unfinished conversation, suitable responses are considered to be contained in similar conversations, retrieved by the ANN model. The last model component receives a retrieved set and ranks possible answers based on answer- and context-relevance. The entire pipeline can be seen in Figure 1.

#### A. Gated Recurrent Unit (GRU)

Gated Recurrent Units (GRUs) are related to Long Short Term Memory Networks (LSTMs) in the sense that they can encode long-term dependencies but have fewer parameters to train and come without an additional cell state. Two gates, the reset and update gates  $r_t$  and  $z_t$ , operate directly on the hidden state, i.e., the hidden layer. Parametrized by  $W$ ,  $U$  and  $b$ , while conditioned on the current input  $x_t$  and previous result  $y_{t-1}$ , GRU gate vectors are computed as:

$$\begin{aligned} z_t &= \varphi_g(W_z x_t + U_z y_{t-1} + b_z) \\ r_t &= \varphi_g(W_r x_t + U_r y_{t-1} + b_r), \end{aligned} \quad (1)$$

with  $\varphi_g$  being the sigmoid function. The update gate  $z_t$  combines the function of the input and forget gate by controlling how much the new hidden state ( $h_t$ ) is defined by either the current input or the last hidden state, using linear interpolation:

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t, \quad (2)$$

with  $\tilde{h}_t$  being the candidate activation. The reset gate  $r_t$  is used to calculate  $\tilde{h}_t$ , controlling similarly how much of the previous hidden state to keep:

$$\tilde{h}_t = \varphi_h(W_h x_t + U_h(r_t \circ h_{t-1} + b_h), \quad (3)$$

with  $\varphi_h$  being the hyperbolic tangent function.

#### B. Hierarchical Recurrent Encoder-Decoder

The HRED model essentially consists of three stacked RNNs: the utterance encoder, context encoder, and utterance decoder, each of them depending on the result of its predecessor and operating on distributed representations.

Formally, a dialogue  $D$ , the input to such a model, can be represented as a sequence of utterances  $D = (U_1, \dots, U_M)$ , with  $U_m$  being a sequence of word indices  $U_m = (w_{m,1}, \dots, w_{m,N_m})$ , each of them usually pointing to a vocabulary reference or directly to a word embedding. These become the input to the utterance encoder.

1) *Encoding steps:* To better capture long-term dependencies, the GRU gating function is used for the individual RNNs. For a simplified notation, the GRU can be expressed compactly by combining equations 1, 2 and 3:

$$h_t = GRU(h_{t-1}, x_t), \quad (4)$$

computing current hidden state  $h_t$ , conditioned on the previous,  $h_{t-1}$  and on current input  $x_t$ . To comply with the HRED notation, the utterance embedding  $h_{m,n}$  of the current utterance  $U_m$ , including word  $w_{m,n}$ , is calculated as:

$$h_{m,n} = GRU_{utt}(h_{m,n-1}, w_{m,n}). \quad (5)$$

Applying equation 5 consecutively on word embeddings  $w_{m,1}, \dots, w_{m,N_m}$ , results in an equally-sized set of hidden states  $h_{m,1}, \dots, h_{m,N_m}$ , where the last hidden state  $h_{m,N_m}$  is the summary of all words in the same utterance. As such, we denote  $h_m = h_{m,N_m}$  to be the hidden state that represents utterance  $U_m$ .

Using this encoding approach, a set of utterances  $U_1, \dots, U_M$  is encoded into hidden states  $h_1, \dots, h_M$ . Those are used as input to the GRU-based context encoder, similar to how word embeddings acted as input to the utterance encoder. As such, context embeddings  $c_m$  are a summary of utterances and represent entire dialogues. They are computed as:

$$c_m = GRU_{con}(c_{m,n-1}, h_m). \quad (6)$$

2) *Decoding step:* In addition to encoding a sequence of embeddings into a hidden state, the decoder component generates word probabilities over a vocabulary, given some context  $U_1, \dots, U_{m-1}$  and previous words  $w_{m,1}, \dots, w_{m,n-1}$ .

Firstly, to condition the decoder RNN on previous utterances, the initialization of its hidden state is based on the context encoders last hidden state  $c_{m-1}$ . If not designed explicitly, context and decoder RNN usually have different hidden state dimensionalities, which is why an additional network layer is added to project context embeddings into the decoder space:

$$d_{m,0} = \tanh(D_0 c_{m-1} + b_0), \quad (7)$$

with parameters  $D_0$  and  $b_0$  and  $d_{m,0}$  being the decoder RNN's initial hidden state.

Given a set of words  $w_{m,1}, \dots, w_{m,n-1}$ , having been previously generated or representing a training example, the decoder RNN hidden state is similarly computed as it was done for the encoder RNNs:

$$d_{m,n} = GRU_{dec}(d_{m,n-1}, w_{m,n}), \quad (8)$$

processing words consecutively. The first iteration uses the hidden state computed by equation 7 and a zero-value embedding for  $w_{m,0}$  to predict the first word of an utterance.

Using both, the hidden state  $d_{m,n-1}$  and word embedding of  $w_{m,n-1}$ , the word embedding of current word  $w_{m,n}$  is then predicted as:

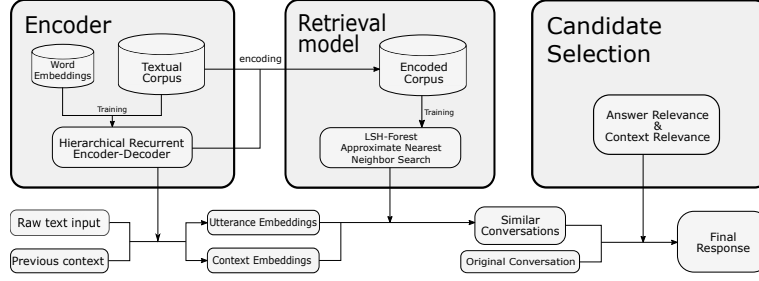


Fig. 1. A view of the pipeline implementing the proposed approach. An HRED encoder is used to generate context and response embeddings and an ANN model builds on previous steps to retrieve similar conversations. Finally, the best candidate is selected according to answer- and context-relevance.

$$w(d_{m,n-1}, w_{m,n-1}) = H_0 d_{m,n-1} + E_0 w_{m,n-1} + b_0, \quad (9)$$

with the additional parameters  $H_0$ ,  $E_0$  and  $b_0$ .  $H_0$  and  $E_0$  control which part of the previous context- and word embedding contribute to the new word embedding and how much of that part is used.

### C. Retrieval Model

Using the encoded corpus as a database of vectors, a Nearest Neighbor Search (NNS) algorithm can be used to find close embeddings in the whole set. For this purpose, an ANN approach has been considered, as general space-partitioning approaches, aiming to improve the NNS performance, suffer from the curse of dimensionality.

Locality Sensitive Hashing (LSH) [11], [12] is an ANN approach that uses a set of hashing functions to project similar data points into buckets and as such, significantly restricts the search space to the size of the bucket. For a projection, a binary string label is constructed by applying  $k$  different hashing functions to a single data point, where the output of such a function is either one or zero. The desired goal of a hashing function is to output the same label for similar data points and differing labels for dissimilar ones. Therefore, binary string labels that are similar, indicate that also the original data points are similar. The string label is then used as a key to index a bucket of similar data points, where a brute force approach can be applied on a much smaller set. A collection of buckets is called a hash-table and  $l$  tables constitute the entire model.

One of the main issues with the basic LSH algorithm [12] is that choosing the optimal number of hashing functions  $k$  and number of tables  $l$  requires one to know the most suitable value for  $r$ , the threshold separating similar and dissimilar points. The LSH-Forest algorithm solves this issue by allowing labels with variable length and thus, eliminating parameter  $k$ .

Instead of linking fixed-length labels to buckets, the label string is stored in a prefix tree, a binary tree (also called 'trie'), in which keys are not contained within nodes but derived from the path that leads from the root to a node.

Each level of the tree is associated with a different hashing function, sampled uniformly and with replacement from a family of hashing functions  $\mathcal{H}$ . Such a tree, an LSH-Tree, is the equivalent to an LSH-based hash-table and the composition of  $l$  trees is an LSH-Forest.

Given a query point  $p$ , finding close neighbors in a set of LSH-Trees  $T_1, T_2, \dots, T_l$  is performed in two phases. First, in a top-down phase or descent, each tree  $T_i$  is searched for the leaf node with the best match to the binary string label of  $q$ .

Inspecting the matches from all trees, the match with the longest prefix defines the tree-level  $x$  from which the bottom-up accumulation, the second phase, begins.

### D. Candidate Selection

Given a query context of an unfinished conversation, using the previously discussed LSH-Forest algorithm, one can retrieve a candidate set from a database of encoded conversations. Candidate answers will be scored based on the matching degree between the retrieved and the original context in terms of question-to-question similarity or answer relevance or other text-based features. However, the scoring functions introduced in this section will solely be based on vector comparison metrics, such as the cosine similarity, as text-based comparison is less rewarding and more difficult and tedious to implement. For the sake of clarity, the query context embedding is defined as  $c_q$ , the textual candidates as  $r_1, r_2, \dots, r_k$ , the context embeddings of candidates as  $c_{r_1}, c_{r_2}, \dots, c_{r_k}$ , and the utterance embeddings of candidates as  $h_{r_1}, h_{r_2}, \dots, h_{r_k}$ .

1) *Context Relevance*: The similarity of two conversations or the distance between a query context  $c_q$  and a candidate context  $c_{r_k}$  has, intuitively, a big impact on the retrieved answer, i.e., the more two questions are similar, the higher the probability that the answers are similar as well. If the cosine similarity has been chosen as the distance function  $D$ , the labels returned by the nearest neighbor search are already sorted by context-to-context distance. Formally, given a candidate response  $r_x$  and a query context  $c_q$ , the Context Relevance (CR) cost function is defined as:

$$\text{cost}_{CR}(r_x) = \text{cos}_{sim}(c_{r_x}, c_q). \quad (10)$$

2) *Answer Relevance*: By manual inspection of near neighbors, it became apparent that the correct answer is usually represented or almost captured in many topic-related candidates. Assuming that the most suitable topic for answering is dominantly represented amongst candidates, responses are ranked based on how much they capture the general topic. Formally, the cost of a response  $r_x$  is defined by the accumulated similarity between its respective embedding  $h_{r_x}$  and the

utterance embeddings of all other candidates (See Figure 2), normalized by length  $k$ :

$$\text{cost}_{AR}(r_x) = \frac{1}{k} \sum_{i=1}^k \text{cos}_{sim}(h_{r_x}, h_{r_i}) \quad (11)$$

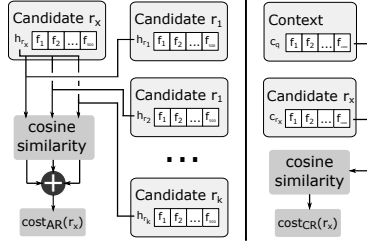


Fig. 2. **Left:** An image showing how the AR cost of a single candidate is accumulated by computing the cosine similarity with other candidate’s utterance embeddings. **Right:** A simple illustration of how CR is computed for a single candidate.

3) *Combining Context and Answer Relevance:* The problem with the previous approach is that the candidates that are off-topic still contribute to the answer relevance cost. Therefore, in a pre-step, according to the previously described context relevance metric, the top  $n$  candidates are accumulated to represent the best general answer topic. In the next step, candidates are ranked based on their similarity to these  $n$  responses. Formally, combined Context and Answer Relevance (CAR) is defined as:

$$\text{cost}_{CAR}(r_x) = \frac{1}{n} \sum_{i=1}^n \text{cos}_{sim}(h_{r_x}, h_{r_i}), \quad (12)$$

with  $n \leq k$ .

#### IV. EXPERIMENTS

##### A. Datasets

The first dataset we use is the Ubuntu Dialogue Corpus which has been studied in most state of the art systems (similar to HRED). The ubuntu dataset contains almost 1 million multi-turn dialogues, with a total of over 7 million utterances and 100 million words. More information can be found in [13].

The second dataset we use is the Vodafone corpus which is created by retrieving archived conversations of the Dutch Vodafone online customer service. Customers having problems with their phone, want to make contractual changes or experience other product related issues, often decide to talk with a Vodafone service agent through an online chat platform.

Every conversation that was not clearly identified as Dutch text was filtered out of the corpus using a port of Google’s Java language detection implementation<sup>1</sup>. Furthermore, to guarantee that the HRED model receives actual conversations for its training, conversations that have less than 5 turns have also been filtered out.

The original corpus contains phone numbers, addresses, names, postal codes and other personal information. To

guarantee anonymization and also to allow enough generalization, this data has been replaced by a meta-token, e.g. "<street\_name>" or "<city>", which is considered to be beneficial for the performance of word embeddings. This way many more training examples will contain these general concepts (like "<street\_name>") and the model can learn in which context a street name should appear. This is possible because the word embeddings of such concepts are also tuned during the training.

The final corpus was generated by using a minimal word occurrence threshold of 10, resulting in a dictionary size of 42892 and an average of 0.435% unknowns per dialogue. The complete statistics for both datasets can be seen in Table I.

	Ubuntu	Vodafone
Language	English	Dutch
Total # of dialogues	487,337	384,897
Total # of turns	2,406,483	6,571,902
Total # of utterances	3,644,566	10,461,677
Total # of words	44,246,198	122,325,433
Avg. # of words per dialogue	90.792	317.81
Avg. # of turns per dialogue	4.938	17.07
Avg. # of words per turn	15.880	18.65
Avg. # of utterances per dialogue	7.479	27.18
Avg. # of words per utterance	11.264	11.58

TABLE I  
STATISTICS OF UBUNTU & VODAFONE CORPUS

##### B. Evaluation process

A quantitative evaluation metric, the Recall@ $k$  measurement [13], has been used to compare the ranking performance of models. Given a context, a set of  $n$  possible answers is presented to a model, which has to rank the answers by their likelihood of being the actual response. For a single evaluation sample, if the correct answer is ranked to be amongst the  $k$  best, the model succeeded. The overall performance of a model is defined as the ratio of correctly ranked answers to all answers, i.e., the percentage of correct answers that were ranked to be amongst the  $k$  best.

By iterating over the conversations in the held-out test set, an evaluation sample has been created for each individual turn or response, with the previous turns representing the context and the current turn or response being the ground truth. In addition to the actual response, a single example also contains  $n - 1$  randomly sampled answers, which the model should preferably rank lower than the true answer.

Each of the models, generative- or retrieval-based, receives the context of a conversation from an evaluation sample and has to generate or retrieve a suitable answer. The utterance-embedding of this answer is then used to compute the distance to each of the 10 possible answers in the evaluation sample, using the cosine similarity between utterance-embeddings. The final ranking is based on this distance, placing similar answers at the top.

As we wanted to have the same conditions for all models, our ranking approach differs from the one used in [13], where answer-embeddings have been directly predicted by an additional network layer. Instead, we used Beam Search (using 5 beams) to generate an answer with the *HRED* model and used the answer’s utterance embedding to compute the ranking for the generative approach.

<sup>1</sup><http://code.google.com/p/language-detection/>

### C. Results and discussion

For the Ubuntu corpus, a HRED model is trained and then the generative approach (of the original model) and the different candidate selection methods (as described in Section III-D) are compared. Results can be found in Table II.

Model	R@1	R@2	R@5
<i>HRED</i>	34.8 $\pm$ 0.4	50.5 $\pm$ 0.4	78.2 $\pm$ 0.3
<i>HRED-CR</i>	32.8 $\pm$ 0.3	47.5 $\pm$ 0.4	74.1 $\pm$ 0.3
<i>HRED-AR</i>	<b>44.1 <math>\pm</math> 0.4</b>	<b>58.6 <math>\pm</math> 0.4</b>	<b>80.5 <math>\pm</math> 0.3</b>
<i>HRED-CAR</i>	43.5 $\pm$ 0.4	58.0 $\pm$ 0.4	80.3 $\pm$ 0.3

TABLE II  
OVERALL RANKING PERFORMANCE OF MODELS ON THE UBUNTU CORPUS. CONFIDENCE INTERVALS ( $\pm 95\%$ ) ARE SHOWN NEXT TO THE AVERAGE PERFORMANCE.

From this Table it is obvious that AR model outperforms other candidate selection techniques as well as the generative approach. When taking into account the context of the whole conversation, results are slightly worse which means that answers are better predicted by focusing on each turn individually rather than taking into account the entire context.

For the Vodafone corpus, three HRED models have been compared, each initialized with a different set of word embeddings in order to assess the effect of local/global context in a language setting other than English. The first model, based only on local domain knowledge, received word embeddings trained with the gensim python library [14], a tool that, given a corpus, will train word embeddings specifically for that corpus. The second model utilized word embeddings from [15], representing global domain knowledge. The embeddings were trained on a corpus consisting of 4 billion words, which was automatically generated by analyzing Dutch websites. The last model received word embeddings that are a combination of the two previously described sets. Both contain word embeddings with a feature-length of 320. However, the embeddings for the last model will have a length of 420, using a concatenation of global embeddings with 320 features and local embeddings with 100 features.

As with the Ubuntu corpus, the generative and retrieval based approaches (*CR*, *AR*, *CAR*) are compared and additionally in this setting, they are also tested upon different embedding initialization approaches (*HRED<sub>L</sub>*, *HRED<sub>G</sub>* and *HRED<sub>LG</sub>*). An overview of the results can be found in Table III.

As expected, for the majority of setups, models can predict assistant responses easier than customer responses. The CAR candidate selection method outperforms all other techniques when predicting assistant responses. However, the performance of customer response prediction is slightly dominated by AR. A reason for this could be that customers often reply with new questions that might not be context related, making answer relevance more important than context relevance.

Furthermore, it can be seen that initializing the HRED model with word embeddings containing global domain knowledge results in the best performance for candidate selection approaches. However, combining global and local domain knowledge has not led to the desired improvements.

This can be explained as follows: The computational graph of the HRED model that defines its training also includes tuning the word embeddings. As such, during the training, the word embeddings are already altered to encode local domain knowledge, even if they were only initialized with embeddings containing global domain knowledge. Adding additional feature-length will in the worst case only add complexity to the model.

The performance of the generative approaches, *HRED<sub>L</sub>*, *HRED<sub>G</sub>*, and *HRED<sub>LG</sub>*, are relatively similar. However, *HRED<sub>L</sub>* slightly outperforms the others. This is contradictory, considering that the candidate selection methods, CR, AR and CAR, clearly perform better on embeddings generated by *HRED<sub>G</sub>* and *HRED<sub>LG</sub>* (See Table III). A reason for this could be that utilizing global domain knowledge to generate an answer is more difficult than using specific domain knowledge. Especially for a very homogeneous (and domain specific) corpus, giving standard answers can work better. Nonetheless, similarity comparisons, used by the NNS approaches, could still benefit from richer embeddings.

Table IV presents some examples of answers using the Ubuntu Corpus using the generative approach (HRED) and the proposed AR model.

### V. CONCLUSION AND FUTURE WORK

End-to-End Dialogue Systems are relatively new and most architectures are far away from being ready for deployment in actual industry that most likely will require more years of research. The architecture proposed in this paper can be seen as a combination of end-to-end and modular Dialogue System. The used retrieval based approach, utilizing dialogue and utterance embeddings which were trained end-to-end, has been shown to outperform the generative approach of the HRED model.

More recently proposed end-to-end systems, the VHRED model and Multiresolution Recurrent Neural Networks [16], both being an improved version of the HRED model, are raising another question: Will one of these models outperform the proposed retrieval based approach, even though all operate on the same embeddings, i.e., at which point does the generative approach benefit from the embeddings' quality more than the retrieval-based? It would be interesting to explore the performance of other encoding models (like the VHRED model and Multiresolution Recurrent Neural Networks). However, one must not underestimate the importance of the proposed ranking system, since it can directly be used by a human agent as a means to assist communication with a client. An interesting research direction arising from this paper would be to allow human agents to affect the ranking score and by this way providing feedback (in terms of reinforcement learning [17]) to the system, which then might be able to re-rank answers.

Another future direction is the simulation of conversations with a tree search, by representing context embeddings as states (tree nodes) and utterance embeddings as actions (connections between nodes). By this way, the search tree can explore possible paths and score responses based on

Approach	Model	Predicting assistant responses			Predicting customer responses		
		R@1	R@2	R@5	R@1	R@2	R@5
Beam Search	<i>HRED<sub>L</sub></i>	<b>31.2 ± 0.9</b>	<b>45.8 ± 0.8</b>	<b>73.4 ± 0.9</b>	28.5 ± 0.7	<b>39.8 ± 0.9</b>	<b>66.1 ± 0.8</b>
	<i>HRED<sub>G</sub></i>	29.9 ± 0.7	44.3 ± 0.9	71.1 ± 0.6	27.1 ± 0.7	37.7 ± 0.9	63.1 ± 0.9
	<i>HRED<sub>LG</sub></i>	30.3 ± 1.0	44.2 ± 0.9	70.6 ± 0.7	<b>28.9 ± 1.1</b>	<b>39.8 ± 1.1</b>	64.6 ± 1.2
Context Relevance	<i>HRED<sub>L-CR</sub></i>	33.4 ± 0.7	48.0 ± 0.8	75.4 ± 0.6	32.8 ± 1.0	47.5 ± 1.0	73.5 ± 0.9
	<i>HRED<sub>G-CR</sub></i>	34.6 ± 1.1	50.0 ± 1.0	76.2 ± 0.7	32.9 ± 0.8	47.3 ± 0.7	74.2 ± 0.8
	<i>HRED<sub>LG-CR</sub></i>	33.9 ± 0.9	48.8 ± 0.8	74.8 ± 0.8	32.5 ± 0.9	48.1 ± 0.7	74.8 ± 0.6
Answer Relevance	<i>HRED<sub>L-AR</sub></i>	39.6 ± 0.7	55.7 ± 0.9	81.1 ± 0.7	40.0 ± 1.1	55.8 ± 1.0	80.4 ± 0.8
	<i>HRED<sub>G-AR</sub></i>	42.7 ± 0.9	58.5 ± 0.8	82.5 ± 0.7	<b>41.0 ± 0.9</b>	<b>56.9 ± 0.7</b>	<b>81.5 ± 0.6</b>
	<i>HRED<sub>LG-AR</sub></i>	43.0 ± 0.8	59.4 ± 0.8	<b>82.7 ± 0.8</b>	40.1 ± 0.9	55.7 ± 0.9	80.0 ± 0.9
Context and Answer Relevance	<i>HRED<sub>L-CAR</sub></i>	41.3 ± 0.8	57.2 ± 0.8	81.2 ± 0.5	39.9 ± 1.0	55.3 ± 1.0	79.6 ± 0.6
	<i>HRED<sub>G-CAR</sub></i>	<b>44.0 ± 0.7</b>	<b>59.8 ± 0.9</b>	82.6 ± 0.7	40.9 ± 0.7	56.8 ± 0.7	80.6 ± 0.7
	<i>HRED<sub>LG-CAR</sub></i>	43.8 ± 0.7	59.5 ± 0.8	82.6 ± 0.7	39.4 ± 0.7	55.1 ± 0.9	79.6 ± 0.6

TABLE III

OVERALL COMPARISON OF MODEL PRECISIONS (IN %). CONFIDENCE INTERVALS (±95%) ARE SHOWN NEXT TO THE AVERAGE PERFORMANCE.

TABLE IV

EXAMPLE ANSWERS OF THE HRED MODEL AND THE RETRIEVAL-BASED APPROACH (USING AR) TO A COUPLE OF QUESTIONS FROM THE UBUNTU CORPUS TEST SET.

Context	Response
I have a netbook __cou__ with 1gb of ram . Will ubuntu 12.04 be fine ? __cou__ With unity of course !	(Original): barely (AR): you don't need much hardware to run gnome , unity perhaps yes though (HRED): that 's a good idea , but it 's not a good idea to stick with it
Hi Guys __cou__ I need urgent help __cou__ I've been trying all day to install Ubuntu on my Mac __cou__ yeah but I can't find ANY solution __cou__ but can someone **unknown** me determining what to do ? __cou__ My **unknown** USB gives me kernel panic __cou__ not syncing	(Original): Why dont you use a cdrom for installing Ubuntu ? (AR): does your Bios support booting usb device ? (HRED): not sure what to tell you , sorry
any help folks. **unknown** __cou__ → help with what ? __cou__ → ; **unknown** hello folks , ok so i will try this again maybe someone here is able to help me with my issue . running 12.04 on an hp **unknown** laptop . when i plug my headphones sound still comes from main speakers and my headphone , how can i make it so that when i plug my headphone the sound will only play thru the headphones	(Original): did you check out the **unknown** link someone posted to you ? (AR): make sure you have the mixer and sound devices installed in your kernel (HRED): that sounds like a hardware issue , right ?
my computer crashed ? how can I see the log ?	(Original): look in /var/log - especially at the xorg log (AR): somewhere in /var/log (HRED): it should be in the menu . lst file
Question : is sshd installed by default with ubuntu 12.04 or do i need to install openssh	(Original): normally installed by default (AR): ubuntu doesn't install an ssh daemon by default so yes (HRED): it should be installed by default
is there a " terminal service " for linux __cou__ i.e. not vnc __cou__ → X ? __cou__ ssh ? __cou__	(Original): i tried it and it was rather laggy (AR): idk it just seems lame lol ; x (HRED): this is what I was looking for
How hard will it be to upgrade from Ubuntu 12.04 to 12.10 ?	(Original): update-manager will prompt you if you want to upgrade when it 's released and you press yes (AR): it is still there , just not installed by default (HRED): first you have to use sudo apt-get dist-upgrade to get it to work

the quality of simulated conversations. Finally, the candidate selection module of our proposed pipeline reveals new opportunities for utilizing such a ranking/similarity model in other problems/domains such as recommender systems. Reviews of products, services, etc. could be encoded using a model like HRED and then based on a query (question) of a user, recommendation can take place by ranking the most relevant reviews (answers).

#### ACKNOWLEDGEMENT

We would like to thank Harry Backers and Marcel Overdijk for their collaboration and support. We gratefully acknowledge the support of QNH Consulting with the donation of the Nvidia GTX 1070 GPU used for this research and for providing the Vodafone dataset.

#### REFERENCES

- [1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [2] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 3776–3783.
- [3] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues," *arXiv preprint arXiv:1605.06069v3*, 2016.
- [4] A. Shawar and E. Atwell, "Chatbots: are they really useful?" in *Journal for Language Technology and Computational Linguistics*, vol. 22, no. 1. GSCS German Society for Computational Linguistics, 2007, pp. 29–49.
- [5] C. D. Manning and H. Schütze, "Foundations of statistical natural language processing," *The MIT Press*, 1999.
- [6] J. Dowding, J. M. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran, "Gemini: A natural language system for spoken-language understanding," in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1993, pp. 54–61.
- [7] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.
- [8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
- [9] P. Shah, D. Hakkani-Tür, and L. Heck, "Interactive reinforcement learning for task-oriented dialogue management," in *NIPS 2016 Deep Learning for Action and Interaction Workshop*, 2016.
- [10] J. Schmidhuber, "Deep learning in neural networks: An overview," in *Neural networks*, vol. 61. Elsevier, 2015, pp. 85–117.
- [11] S. Har-Peled, P. Indyk, and R. Motwani, "Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality," in *Theory of computing*, vol. 8, no. 1, 2012, pp. 321–350.
- [12] M. Bawa, T. Condie, and P. Ganesan, "LSH Forest: self-tuning indexes for similarity search," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 651–660.
- [13] R. Lowe, N. Pow, I. V. Serban, and J. Pineau, "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems," in *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015.
- [14] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, <http://is.muni.cz/publication/884893/en>.
- [15] S. Tulkens, C. Emmery, and W. Daelemans, "Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odiijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), may 2016.
- [16] I. V. Serban, T. Klinger, G. Tesauro, K. Talamadupula, B. Zhou, Y. Bengio, and A. Courville, "Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation," *arXiv preprint arXiv:1606.00776*, 2016.
- [17] J. D. Williams and G. Zweig, "End-to-end lstm-based dialog control optimized with supervised and reinforcement learning," *arXiv preprint arXiv:1606.01269*, 2016.